УДК 004.023

В. В. Гаращенко
Научный руководитель: канд. тех. наук О. В. Желудкова
Тюменский государственный университет, Тюмень

АЛГОРИТМ СКРЫТИЯ КОНФИДЕНЦИАЛЬНЫХ АССОЦИАТИВНЫХ ПРАВИЛ

Аннотация. В связи с необходимостью защиты не только данных, но и сведений, полученных в результате их анализа, в данной работе рассмотрена проблема скрытия ассоциативных правил. Предложена модификация алгоритма скрытия конфиденциальных ассоциативных правил, позволяющая в некоторых случаях уменьшить количество потерянных и ложных правил.

Ключевые слова: информация; конфиденциальность; Data Mining; ассоциативные правила.

Поиск ассоциативных правил является одной из задач интеллектуального анализа данных. Его целью является нахождение зависимостей между связанными событиями или элементами.

В настоящее время распространена практика обмена данными между организациями в процессе делового сотрудничества. Аналитиками могут использоваться различные методы интеллектуального анализа данных для извлечения полезной информации из данных, полученных от партнеров. Однако, несмотря на все преимущества для бизнеса, такой анализ может представлять угрозу раскрытия конфиденциальных сведений посторонним лицам, например критически важной информации, обеспечивающей конкурентное преимущество организации. В таком случае существует необходимость защиты не только самой информации, но также и некоторых данных, полученных в результате анализа. Поэтому возникает вопрос: как сохранить конфиденциальность данных при условии необходимости обмена информацией или ее публикации и последующего анализа?

Эта проблема бросает вызов традиционному анализу данных, поэтому для решения такого рода задач широко используется интеллектуальный анализ данных, сохраняющий конфиденциальность (Privacy Preserving Data Mining). В данной работе рассмотрено скрытие ассоциативных правил.

При поиске ассоциативных правил используют показатели, позволяющие определить их значимость. Основными показателями являются поддержка и достоверность. Поддержка правила $X \Rightarrow Y$ — величина, показывающая, какая доля транзакций содержит одновременно наборы X и Y. Достоверность прави-

ла показывает, какова вероятность того, что вместе с набором X в транзакции встретится набор Y. Для ограничения количества найденных правил при поиске задаются минимальные пороговые значения поддержки (MST — minimum support threshold) и достоверности (MCT — minimum confidence threshold) [1].

Проблему скрытия конфиденциальных ассоциативных правил можно выразить следующим образом. Пусть D — база данных транзакций и R — множество правил, которые могут быть найдены в D при заданных значениях MST и MCT. Пусть R_s — множество правил, которые необходимо скрыть, $R_s \subset R$. R_n — множество остальных правил, $R_n \cup R_s = R$. Процесс скрытия заключается в преобразовании D в базу данных D', в которой в процессе поиска ассоциативных правил могут быть обнаружены лишь правила из R_n. $|D| = |D'|$. R' — множество правил, полученных из D' с теми же значениями MST и MCT.

Пусть Δ — набор транзакций, которые должны быть преобразованы в D для того, чтобы скрыть правила R_s. Тогда $\Delta \subseteq D$ и $\Delta = D - D'$. $|\Delta|$ — количество преобразованных транзакций. Пусть Δ' — транзакции, которые привели к изменениям в D'. Тогда $\Delta' \subseteq D'$ и $\Delta' = D' - D$. $|\Delta| = |\Delta'|$. Цель скрытия заключается в поиске набора Δ и преобразований Δ в Δ' [2].

Таким образом, для скрытия ассоциативных правил осуществляется преобразование исходного набора данных. Это в свою очередь ведет к побочным эффектам: при поиске правил в преобразованном наборе с теми же значениями MST и MCT могут быть потеряны некоторые правила, которые изначально могли быть найдены в исходном наборе данных (lost rules), а также могут появиться новые, ложные правила (ghost rules). Поэтому исследователям приходится искать компромисс между конфиденциальностью и точностью данных.

Выделяют три основных подхода к скрытию ассоциативных правил: эвристический, граничный (border-based) и точный (exact). Первый подход является наиболее распространенным среди исследователей. Методы эвристического подхода включают возмущение, заключающееся в удалении/добавлении элементов или добавлении шума, и блокирование, основанное на замене известных значений неизвестными, например «?» [3].

Различные эвристические алгоритмы используют разные стратегии по выбору транзакций и элементов для преобразования набора данных. Это два основных аспекта, влияющих на эффективность алгоритмов. Основной целью разработки новых алгоритмов является уменьшение количества нескрытых конфиденциальных правил и уменьшение нежелательных побочных эффектов, таких как потерянные и ложные правила.

В своей работе мы опирались на статью [4]. В ней представлен эвристический алгоритм HSARWI, который скрывает все выбранные конфиденциальные ассоциативные правила. Как и большинство существующих алгоритмов, HSARWI скрывает ассоциативные правила путем уменьшения их поддержки

и достоверности. Достигается это за счет удаления элементов из выбранных транзакций.

В алгоритме для каждой транзакции и каждого элемента в транзакции вычисляется вес (WT_i и WI_{ik} соответственно). Для преобразования базы данных транзакций до тех пор, пока не останется нескрытых конфиденциальных правил, выбирается транзакция с наибольшим весом WT_i, и в ней удаляется элемент с наибольшим весом WI_{ik}. Если несколько элементов имеют одинаковый вес, элемент для удаления выбирается случайным образом. В работе [4] приведен псевдокод алгоритма.

Для того чтобы улучшить алгоритм, а именно сократить число потерянных и ложных правил, так как HSARWI гарантированно скрывает все конфиденциальные правила, было добавлено условие: если несколько элементов имеют одинаковый вес, кандидата для удаления выбирать не случайным образом, а выбирать тот, который реже встречается в неконфиденциальных правилах. Это позволит добиться того, что удаление элемента приведет к изменению поддержки и достоверности в меньшем количестве неконфиденциальных правил.

Алгоритм HSARWI и его модификация были реализованы на языке Python. Для поиска ассоциативных правил использовалась библиотека apyori, для сравнения использовались базы данных Chess и Mushroom. Результаты показали, что данное изменение в алгоритме позволило в некоторых случаях уменьшить количество потерянных и ложных правил, в большинстве же случаев результат остался прежним. На рис. 1 приведены некоторые примеры сравнения алгоритма HSARWI и его модификации.

Конфиденциальные правила	HSARWI		Модификация	
	ghost	lost	ghost	lost
39, 90 - >36 85, 90 -> 39	58	419	30	337
24, 34, 90 -> 85 1, 34, 86 -> 85 34, 59 -> 36	9	619	9	603
24, 36, 85 -> 34 24, 86 -> 1 63 ->36 2, 85, 86 -> 39	12	541	12	473

Mushroom, MST = 0.4, MCT = 0.6

Конфиденциальные правила	HSARWI		Модификация	
	ghost	lost	ghost	lost
7, 58 -> 60 36, 60 -> 29	0	308	0	245
29, 62 -> 58 7, 29, 60 -> 52 29, 36, 40 -> 60	0	336	0	311
29, 36, 52, 58 -> 40 52, 58, 60, 66 -> 29 34, 52, 58 -> 40 36, 40, 52, 60 -> 29	0	221	0	149

Chess, MST = 0.93, MCT = 0.97

Рис. 1. Сравнение результатов работы алгоритмов

Планируется дальнейшая работа над алгоритмом, а именно над функцией веса элемента, чтобы попытаться значительнее уменьшить количество потерянных правил. Также будет разработано web-приложение, включающее модули поиска ассоциативных правил, скрытия ассоциативных правил и сравнения результатов для удобной работы.

Список литературы

1. Методы поиска ассоциативных правил [Электронный ресурс]. URL: http://www.intuit.ru/studies/courses/6/6/lecture/186 (дата обращения: 21.10.2017).
2. *Cheng P.*, *Lin C.-W.*, *Pan J.-S.* Use HypE to hide association rules by adding items. Shenzhen, 2015.
3. *Sathiyapriya K.*, *Sudha Sadasivam Dr. G.* A survey on privacy preserving association rule mining. Coimbatore, 2013.
4. *Sakenian Dehkordi M.*, *Naderi Dehkordi M.* Introducing an algorithm for use to hide sensitive association rules through perturb technique. Isfahan, 2016.

УДК 004.056

Р. В. Гибилинда
Научный руководитель: канд. тех. наук Д. А. Хорьков
Уральский федеральный университет, Екатеринбург

СПОСОБ ВОССТАНОВЛЕНИЯ ДАННЫХ В ФАЙЛОВОЙ СИСТЕМЕ EXT4 С ИСПОЛЬЗОВАНИЕМ ИНФОРМАЦИИ ЖУРНАЛА ИЗМЕНЕНИЙ ТОМА

Аннотация. В докладе рассмотрена проблема восстановления файлов в файловой системе ext4, указаны ее причины. Описан формат файла журнала изменений тома, позволяющего поддерживать раздел в исправном состоянии. Продемонстрирован способ, позволяющий как в ручном, так и в автоматическом режиме производить восстановление данных на разделе без использования сигнатурного поиска. Указаны достоинства и недостатки предложенного способа.

Ключевые слова: информация; восстановление; журналирование; доступность информации; ext4.

Основная сложность восстановления информации в ext4 связана с тем, что при удалении последней жесткой ссылки на файл драйвер заполняет нулями область индексного узла, где указаны номера кластеров с данными файл [1]. Анализ битовой карты с целью последующего исследования свободных областей памяти на машинном носителе занимает продолжительное время, особенно на накопителях большого объема и RAID-массивах. Для ускорения процесса поиска данных удаленного файлового объекта требуется возможность быстрого обнаружения номеров его кластеров. Ее предоставляет журнал изменений тома — файл, призванный обеспечить отказоустойчивость раздела с точки зрения его логической структуры. Формат журнала и алгоритм его работы не